DOCUMENT RESUME

ED 052 247 TM 000 653

AUTHOR	Boldt, Robert F.
TITLE	Comparability of Scores from Different Tests Though on the Same Scale.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-71-10
PUB DATE	Feb 71
NOTE	13p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Comparative Analysis, *Comparative Testing, *Measurement Techniques, Norms, Psychological Tests, *Scores, *Scoring, *Testing, Tests
IDENTIFIERS	*Scaled Scores

ABSTRACT

Scores from tests in the same battery are put on scales which are the "same" in some sense, so that certain interpretations are made easier. This is often done when scores for different tests are obtained on different population segments, especially such as newer, more varied batteries of test offerings. It is felt that traditional erroneous expectations about the meaning of scaled scores may be carried over into the new situations and hence certain of these expectations are discussed. It is suggested that when special properties of scales are considered valuable for the users of a new battery, active technical steps beyond those of traditional scaling are required to assure that these values are implemented. (Author)

ED052247

RESEARCH BULLETIN

RB-71-10

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

COMPARABILITY OF SCORES FROM DIFFERENT TESTS THOUGH ON THE SAME SCALE

Robert F. Boldt

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
February 1971

TM 000 653

Comparability of Scores from Different Tests

Though on the Same Scale

Abstract

Scores from tests in the same battery are put on scales which are the "same" in some sense, so that certain interpretations are made easier. This is often done when scores for different tests are obtained on different population segments, especially such as newer, more varied batteries of test offerings. It is felt that traditional erroneous expectations about the meaning of scaled scores may be carried over into the new situations and hence certain of these expectations are discussed. It is suggested that when special properties of scales are considered valuable for the users of a new battery, active technical steps beyond those of traditional scaling are required to assure that these values are implemented.

Comparability of Scores from Different Tests Though on the Same Scale

R. F. Boldt

It is customary for those who offer psychological tests to follow the very useful practice of reporting (or recommending) transformed raw (or formula) scores so that the reported scores will be on a system of numbers which is familiar in some sense. By far the most common method used is to set the mean and standard deviation equal to some handy integral values such as 500 and 100 used by the Educational Testing Service, 100 and 20 for tests used by the Army, or 50 and 10 used by the Navy. The choice of mean and standard deviation is for convenience and unambiguity rather than in response to technical requirements, and these means and standard deviations themselves apply to particular populations on which the tests were standardized, e.g. the Army uses the World War II mobilization population while for the Scholastic Aptitude Test (SAT) the College Entrance Examination Board (CEEB) uses the group of all students who took the SAT in April 1941. The person who achieves a verbal score of 500 on the Graduate Record Examination (GRE) is not equivalent to a person who achieves a score of 500 on the SAT because the reference groups are quite different.

The establishment of scales for a battery of tests presents a more complex problem when one wants the tests of the battery to be "on the same scale" in some sense. The sense in which this scaling is to be accomplished depends on the use to which the tests are to be put. One such use is that of the military service (Wolfe, 1969) where a large aggregation of people are to be examined to reach an acceptance-type decision (induction), or to be sorted into a variety of categories the members of which are later to receive

specialized training of some kind. In this case all people take the same tests and, except for the test results, can all be considered candidates for the same sets of activities. Modern trends in educational philosophy, however, push the test publisher toward a more flexible type of offering where some set of measures is administered to all candidates (core tests) but others are offered to or chosen by particular candidates according to the needs of the situation. The trend seems to be away from testing as a selection hurdle, and toward testing as a way for the candidate to demonstrate something he can do. Different people would choose to exhibit different aspects of themselves and hence similar decisions would not necessarily be based on the same information. Scholarship application is one traditional context for such a situation, and so is application to a professional school. In both cases an admissions-type (i.e., dichotomous accept-reject) decision is to be made about people for whom nonuniform information is available. Some of the items of information are available on large numbers (though not all) of candidates and it is desired to so score these different kinds of information so that the numbers are comparable in at least some limited sense.

If the population which is to act as the reference group is available for use or for sampling, then one can administer the test to be scaled and use a linear transformation of raw (or formula) scores which yields a prespecified mean and standard deviation. Where only partial information is available on a selected population, however, the test may still be scaled to produce a prespecified mean and standard deviation in a suitable reference group provided one is willing to behave as if the selected population differs from the reference population in ways that may be accounted for by explicit

selection on variables measured in both populations. This assumption, when applied to the commonly assumed linear and homoscedastic systems of variables, results in the equations for accounting for the effects of selection as given, for example, in Gulliksen (1950, Ch. 13). The equations given by Schultz and Angoff (1956) derive from the same equations and were introduced initially by L. R Tucker to achieve the scaling of test batteries like the CEEB system of achievement tests or the GRE system of advanced tests. In both systems, the verbal and quantitative aptitude scores take the role of explicit selectors. A particular CEEB Achievement Test or GRE Advanced Test is, of course, taken only by those candidates who choose to take it, so that the same information is not available on all candidates. In fact, for psychological reasons it is clearly meaningless, particularly in the case of GRE advanced tests, to talk about the distribution of advanced test scores as if each of the advanced tests are administered to a single group. One can, nevertheless, think of a series of candidate reference populations, the members of which have backgrounds suitable for the advanced area in question and whose aptitude scores have whatever joint distribution is ascribed to the reference groups. If the advanced tests are scaled so that they have the desired mean and standard deviation in these reference groups then the advanced test scores are comparable (in this particular restricted sense).

Comparability in the restricted sense achieved as described above is quite difficult to express in plain words as one can see by examining contemporary explanations of score scales (or by reading the explanation above). Such explanations lack the clarity of a technically precise expression to a technician, and the simplifications in language attempted for more general audiences do not seem to achieve the bringing of truth to laymen. Hence misunderstandings

arise, two of which will now be discussed. The first of these is that people with equal advanced test scores, though in different areas, will perform the same in subsequent course work. A variant of this is that if a person receives a score in some area and then spends an equal amount of time preparing for a second area as he did for the first, he can be expected to achieve the same score in the second area. Either of the statements may be true about particular individuals but neither is probably true by and large. There is nothing in test construction and scaling technology as described above that suggests that people with equal scaled scores though in different areas of study are interchangeable--nothing is done to bring this interchangeability about; nothing quantitative and lawful is known to indicate that it will come about. The relative size of the scores has meaning only with respect to different reference groups (Smith is higher in his reference group than is Jones in the other reference group).

A second misinterpretation is that the scaling procedure which uses the aptitude scores to achieve comparability leads to a scaling where equal aptitude scores imply equal advanced scores on the average. That this is a misinterpretation can be seen as follows: Suppose that the advanced tests have been scaled so that their standard deviations are equal in their respective reference populations and that only one core variable is used (only one test is administered to all people and can be used to achieve the scaling). The reference populations are assumed to be so composed that the standard deviations of the core variable are the same and that the scaling has been set up so that the standard deviations of the advanced tests in the reference population are equal to that of the core variable. Then under the assumptions given in Gulliksen all the coefficients of regression of the advanced tests on

the core variables would be equal both on the populations tested and on the reference populations. Such a result does not hold across the varying contents of the GRE, for example, since one can see in Table 1 that the regression coefficients of advanced tests on verbal or quantitative tests are not equal in the report of Schultz and Angoff (1956), nor are they equal in a more recent but similar study by Wallmark (1969). In both of these studies the tests have been scaled in the first sense and the tables indicate that the results are not compatible with an interpretation of comparability in the present sense.

Table 1

Univariate Weights of Regressions of Advanced Tests (y) on the GRE Aptitude Tests (V^a, Q^b)

Area	1952			1969		
	N	b_{yV}	b_{yQ}	N	b_{yV}	b_{yQ}
Biology	209	.51	.64	4,696	.64	.61
Chemistry	18	.40	.54	2,416	.43	.64
Economics	239	.62	.50	1,930	.62	.67
Education	180	.80	.56	2,746	.74	.44
Engineering	151	.62	.55	4,259	.37	.77
French	32	.54	.13	1,292	.67	.40
Geography				306	.54	.51
Geology	35	.60	.59	575	.47	.55
German	10	.00	-.34			
Government	146	.72	.49			
History	181	.64	.46	4,919	.58	.38
Literature	239	.83	.52	6,276	.74	.37
Mathematics	81	.21	.50	3,279	.45	.78
Music				647	.57	.51
Philosophy	31	.76	.24	793	.69	.44
Physics	49	.59	.56	2,190	.36	.71
Political Science				2,745	.63	.47
Psychology	171	.61	.51	5,643	.67	.48
Sociology	127	.69	.66	2,151	.76	.63
Spanish	34	.30	.14	770	.35	.17
Speech				695	.58	.36

[a] V for verbal aptitude.

[b] Q for quantitative aptitude.

Hence, by and large one would expect that if the standard deviations of advanced tests are set equal in reference populations which are assumed to have equal standard deviations on the core variable, then it will not be true that equal core variable scores imply equal advanced test scores. In the special case where the standard deviations of the core variable equal that of the advanced test in the reference populations, the average score implied by the aptitude variable would equal the scores on the core variable only if the regression coefficient, and hence the correlation coefficient, equals unity--a result that one is simply not going to experience. Parenthetically, one may note in Table 1 that the regression coefficients are not of the same relative magnitude across all areas so that using V as a core variable would yield a different result than if Q were used as a core variable.

Next consider the case of two core variables. This case is more common since verbal and quantitative tests are often used as core variables in test populations. Suppose that the reference populations are defined with means, standard deviations, and correlations (500, 100, and .4 in the CEEB and GRE systems). We suppose that equal scores on the core variables V and Q imply possibly different scores on the advanced tests but a transformation of the core test scores is sought which would result in equal advanced test scores on the average. Let

$$a_1V + b_1Q + c_1$$

and

$$a_2V + b_2Q + c_2$$

be the regression functions for areas one and two, respectively, where V and Q are the core tests, where the advanced test scores are the dependent

variables. By scaling the advanced tests, one would in effect transform the regression functions to get

$$G_1(a_1V + b_1Q + c_1) + H_1 = F_1(V,Q)$$

and

$$G_2(a_2V + b_2Q + c_2) + H_2 = F_2(V,Q) \quad .$$

If the G's and H's can be chosen so that the F's are the same for all values of V and Q, then we have a system where equal values of V and Q imply equal advanced scores on the average (since these are regression equations). Setting the F's and the resulting coefficients of V and Q equal, one obtains

$$G_1a_1 = G_2a_2$$

$$G_1b_1 = G_2b_2$$

and

$$G_1c_1 + H_1 = G_2c_2 + H_2 \quad .$$

Clearly, given G's one can find suitable H's, using the third equation. Solving for G_2 in terms of G_1 and substituting in the second equation gives

$$G_1b_1 = G_1(a_1/a_2)b_2$$

or

$$b_1/b_2 = a_1/a_2 \quad .$$

In words, if the regression coefficients are proportional, the solution is that the G's and H's are proportional to the ratios of the a's and b's. These results are quite restrictive on the regression equations, so much so that they suggest that even in a battery with as limited a variety

of tests as are available in the GRE and CEEB offerings, the relations among advanced test scores and core variables differ enough to preclude a scaling of the sort sought. Table 2 shows the regression coefficients and their ratios for the Schultz-Angoff and the Wallmark data. In this table the ratios of regression coefficients group the variables in ways which are intuitively reasonable. To digress, one can note the shift in the relative size of the quantitative weight in the fields of economics, physics, engineering, and mathematics. The first three of these has had an increasing emphasis on mathematical analysis in recent years and mathematics itself has tended to

Table 2

Weights and Ratios of Weights of Bivariate Regressions
of Advanced Tests on GRE Aptitude Tests

Area	1952			1969		
	$b_{yV.Q}$	$b_{yQ.V}$	$b_{yQ.V}/b_{yV.Q}$	$b_{yV.Q}$	$b_{yQ.V}$	$b_{yQ.V}/b_{yV.Q}$
Biology	.30	.49	1.64	.46	.34	.73
Chemistry	.12	.47	3.90	.20	.47	2.34
Economics	.47	.31	.66	.39	.40	1.02
Education	.73	.13	.18	.80	.11	.22
Engineering	.48	.26	.55	.19	.69	3.57
French	.71	-.38	-.53	.55	.02	.04
Geography				.31	.30	1.00
Geology	.43	.35	.81	.29	.34	1.17
German	.09	-.33	-3.48			
Government	.64	.16	.27			
History	.58	.13	1.22	.41	.06	.15
Literature	.80	.07	.09	.64	-.02	-.03
Mathematics	.08	.47	6.24	.32	.93	2.94
Music				.34	.23	.69
Philosophy	.85	-.21	-.25	.67	.19	.28
Physics	.51	.22	.44	.27	.82	3.04
Political Science				.45	.14	.32
Psychology	.51	.30	.60	.49	.15	.32
Sociology	.56	.30	.54	.58	.26	.45
Spanish	.40	-.18	-.45	.39	-.05	-.14
Speech				.44	.10	.22

deemphasize analysis. The quantitative tests currently in use are all mainly about analysis from a mathematical point of view and hence mathematics looks rather more verbal than it once did.

These differences in the relations of the weights do not allow one to scale all the advanced tests so that equal aptitude scores imply equal scores on the advanced tests on the average, but subsets of areas might be chosen where the proportionality holds or holds approximately. For example, the comparability among fields may be needed to choose among candidates for some award or acceptance decision which would draw people from fields where the coefficients of regression of advanced test scores on aptitude tests happen to be in about the same ratio. The decision to make science awards at the graduate level to selected students from physics and engineering curricula would be one where comparability is desired and might be achieved, approximately at least, since in the Wallmark data the ratios of regression coefficients are about three-to-one. However, if the biological or social sciences were involved, the comparability could not be achieved. In that case normative relations within fields could be retained but comparisons across fields would reflect to a considerable extent the average aptitude level of people entering the fields.

Sometimes users of test scores have hard decisions to make about awarding things or admitting people. Help is needed in these decisions and users of the scores may understandably want some comparative scores to bolster choices made. The philosophy of value that underlies the production of numbers to be compared is the philosophy of value that makes decisions about people when the numbers are used. When the decisions are made about people using different information, the scaling of the information effects the decision process and

hence implements the values of that process. But the values implemented by the scaling process may not be, and probably are not, those of the user. The two systems of values are not necessarily in conflict, rather they are probably unrelated. Therefore, if the user wishes his values to carry the most emphasis, he must undertake active technical steps to assure that the process that produces numbers or scores to be compared incorporates his particular interests. This is probably not accomplished with existing techniques.

References

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Schultz, M. K., & Angoff, W. A. The development of new scales for the aptitude and advanced tests of the Graduate Record Examinations. Journal of Educational Psychology, 1956, 47, 285-294.

Wallmark, M. M. A rescaling study of the Graduate Record Examinations Advanced Tests (Based on national program data for 1967-68). Educational Testing Service, Statistical Report 69-4. Princeton, N. J., 1969.

Wolfe, J. H. A review of Rulon, P. J., et al. "Multivariate Statistics for Personnel Classification." Educational and Psychological Measurement, 1969, 29, 541-544.